Open Data Policy--
Managing Information as an Asset

Executive Order 13642 &
OMB Memorandum M-13-13
May 9, 2013

He read it so you don't have to!

v2 06-10-2013

Origin and provenance

- Latest in a series of policy directives mandating, among other things, broader public access to Federal and Federally-supported data and information:
  - Transparency and Open Government -- Presidential Memorandum, 2009
  - Open Government Directive -- OMB Memorandum, 2009
  - Digital Government: Building a 21st Century Platform to Better Serve the American People -- Federal CIO Strategy Document, 2012
  - Managing Government Records Directive -- OMB-NARA Memorandum, 2012
  - *Increasing Access to the Results of Federally Funded Scientific Research* -- OSTP Memorandum, Feb. 2013
  - *Making Open and Machine Readable the New Default for Government Information* -- Executive Order, May 2013
  - *Open Data Policy--Managing Information as an Asset* -- OMB Memorandum, May 2013

Motivation, intent, and scope

- Crux of the policy (Executive Order):

  "To promote continued job growth, Government efficiency, and the social good that can be gained from opening Government data to the public, the *default state of new and modernized Government information resources shall be open and machine readable*.

  Government information shall be managed as an asset throughout its life cycle to promote interoperability and openness, and, wherever possible and legally permissible, to ensure that data are released to the public in ways that make the data easy to find, accessible, and usable."

- Applies to ALL Federal data, not just scientific
  - Science-specific requirements addressed in the OSTP's Feb. 2013 memorandum on "*Increasing Access to the Results of Federally Funded Scientific Research*"

Basic desires and (hoped for) effects

- A framework to help institutionalize principles of *effective information management* at each stage of the information's life cycle
  - To promote interoperability and openness
- Commitment to "collect or create information in a way that supports downstream information processing and dissemination activities"
  - Including using machine-readable & open formats, data standards, and common core & extensible metadata for all new information creation and collection efforts
- Ensure information stewardship (and availability for re-use) through the use of open licenses
  - While respecting privacy, confidentiality, security, or other restrictions to release
- Building or modernizing information systems in a way that
  - Maximizes interoperability and information accessibility
  - Maintains internal and external data asset inventories
  - Enhances information safeguards
  - Clarifies information management responsibilities.

Open data defined

- Per OMB M-13-13, 'open data' is "... publicly available data structured in a way that enables the data to be fully *discoverable* and *usable* by end users."
- General attributes:
  - **Public** -- '*presumption in favor of openness*' (with caveats for PII, security, etc.)
  - **Accessible** -- 'convenient, modifiable, and *open formats* that can be retrieved, downloaded, indexed, and searched'
  - **Described** -- '*robust, granular metadata*' & 'thorough description of data elements, data dictionaries', etc.
  - **Reusable** -- 'made *available under an open license* with no restrictions on use'
  - **Complete** -- 'published in *primary forms*' ... *derived* or *aggregate data 'must reference primary data*' (i.e., must document *provenance*)
  - **Timely** -- 'made available as quickly as possible to *preserve the value* of the data'
  - **Managed** post-release -- 'point of contact ... to assist with data use ... respond to complaints about adherence to policy requirements' (i.e., '*user support*')

Scope of the policy

- Applies to:
  - All *new* information collection, creation, and system development efforts
  - Major modernization projects that update or redesign existing information systems
  - "Subject to the availability of funding"
- For existing datasets:
  - Agencies encouraged to "improve the discoverability and usability of existing datasets"
  - Prioritize datasets with emphasis on those
    - Previously release to the public
    - Deemed high-value or high-demand
  - Consider cost/benefit of retrospective documentation, processing, and/or release

Policy requirements (III.1)

- Collect or create information in a way that supports downstream information processing and dissemination activities
  - *Machine-readable data collection default*
    - Use machine-readable and open formats when possible
      - Non-proprietary, publicly available, non-restrictive
    - Implied need to convert (e.g., phone, paper surveys) to digital
  - Use *data standards* -- accepted/acceptable -- not prescribed
    - Promote interoperability
  - Use *open licenses* (as permitted)
    - "grant permission to access, re-use, and redistribute a work with few or no restrictions"
  - Use *common core and extensible metadata*
    - Common core = Data.gov schema
    - Extensible = FGDC, ISO-19139, NIEM, controlled vocabularies, etc.

Policy requirements (III.2)

- Build (or modernize) information systems to support interoperability and information accessibility
- Attributes:
  - Scalable, flexible design to facilitate extraction of data in multiple formats and for a range of uses as internal and external needs change
  - Data outputs must meet all requirements for effective processing and dissemination (e.g., structure, description) and be recorded in the *mandated harvestable data catalog*
  - Data schema and dictionaries are documented and made available to partners and the public

Policy requirements (III.3)

- Strengthen data management and release practices
  - Create and maintain an internal *enterprise data inventory*
    - Eventually to include *all* agency datasets "to extent practicable"
    - Identify data that are already public or *can be made public*
    - As for the public catalog, describe using common core & extensible metadata
  - Create and maintain a *consolidated, harvestable public data listing* (catalog)
    - Includes data generated by agency & through funded grants, cooperative agreements, etc., (excluding administrivia) and
    - Datasets that *can be made public but have not yet been released*
  - Create a process to engage with customers to help facilitate and prioritize data release
  - Clarify roles and responsibilities for promoting efficient and effective data release practices

Harvestable public data catalog

- http://project-open-data.github.io/catalog/
  - All agency data that *are* or *can be* made available
  - Described at a minimum with common core metadata
  - Machine and human-readable
- Machine-readable catalog
  - JSON encoding mandated for interchange of "raw" (non-geospatial) records
  - RDFa Lite and XML optional for additional attribute level metadata
- Apply "standard citation information, preferably in the form of a persistent identifier" to datasets "where feasible"
- Employ controlled vocabularies & *folksonomies* to aid discovery
- Consistently placed public 'web folder' -- *www.agency.gov/data*
  - Facilitate automatic aggregation by Data.gov, et al.

Policy requirements (III.3)

- Strengthen data management and release practices (continued)
  - Create a process to engage with customers to help facilitate and prioritize data release
    - Identification of priority datasets
    - Methods of release
      - Bulk download
      - New APIs
  - Clarify roles and responsibilities for promoting efficient and effective data release practices
    - Internal communication, coordination
      - Work with Privacy and Security officials to assess and minimize risk of release (legal, security, etc.)
    - Engage with "entrepreneurs and innovators" to promote new data uses, applications, and services
      - Challenge.gov "Challenge Yourself to App-lify USGS Data"

Policy requirements (III.4 & 5)

- Strengthen measures to ensure that privacy and confidentiality are fully protected and that data are properly secured
  - Presumption "in favor of openness"
    - Determination to withhold release (based on security, privacy, contractual, or other criteria) must be documented
  - Consider "mosaic effect" of data aggregation
- Incorporate new interoperability and openness requirements into core agency processes
  - Agencies must describe how they have "institutionalized and operationalized" the interoperability and openness requirements "into their core processes across all applicable agency programs and stakeholders"
  - Lots of annual & quarterly reporting and tracking:
    - IRM Strategic Plan, GPRA, growth of catalogs, etc.

Required actions & deadlines

- Six (count them, 6) months from date of Memorandum:
  - Agencies and interagency groups must review and revise existing data management and release policies and procedures to conform to the Open Data Policy
- Agencies must:
  - *Create* and maintain the internal *enterprise data inventory*
  - *Create* and maintain the *public data listing (catalog)*
  - Create a customer engagement process
  - Clarify roles & responsibilities for promoting efficient and effective data release
- Other deadlines to:
  - Make changes to Federal acquisition & grant-making processes
  - Start progress reporting
  - Publish government-wide "open online repository of tools and best practices"

Resources: *Project Open Data*

- An "online repository of tools, best practices, and schema to help agencies adopt the framework presented in this guidance".
  - Reference site for all aspects of the open data policy
  - Editable (requires GitHub registration & approval of content)
  - http://project-open-data.github.io/
- Implementation Guide
  - http://project-open-data.github.io/implementation-guide/
- Open Data Catalog (www.agency.gov/data) guidance
  - http://project-open-data.github.io/catalog/
- GitHub code repository
  - https://github.com/project-open-data

Policy references

- Transparency and Open Government: Presidential Memorandum (01-21-2009)
- Open Government Directive: OMB Memorandum M–10–06 (12-08-2009)
- Digital Government: Building a 21st Century Platform to Better Serve the American People: Federal CIO Strategy Document (05-23-2012)
- Managing Government Records Directive: OMB-NARA Memorandum M–12–18 (08/24/2012)
- Increasing Access to the Results of Federally Funded Scientific Research: OSTP Memorandum (02-22-2013)
- Making Open and Machine Readable the New Default for Government Information: Executive Order 13642 (05-09-2013)
- Open Data Policy-Managing Information as an Asset: OMB Memorandum M–13-13 (05-09-2013)

Questions?

None? Good, let's get on with the discussion!

Open discussion-- suggested topics

- How might these requirements impact USGS? For example:
  - Need for modification, expansion, or better integration of existing data management & delivery applications, systems, or services ?
  - Need for altered workflows and data management responsibilities within/among mission areas ?
  - Need for additional access control methods, e.g., on pre-release data entries in the mandated public data catalog ?
- Relation to recent OSTP research access memorandum:
  - Implications for documenting, preserving, managing, cataloging, linking/relating, and serving:
    - 'Primary' data & data 'disentangled' from publications
    - Code, methods (e.g., models, workflows & procedures), physical objects (e.g., samples)
- Mandated agency harvestable data catalog:
  - Role of USGS Core Science Metadata Clearinghouse ? ScienceBase ? Catalog of USGS Data ?
  - Other internal or external systems or services ?

What do data consumers want?

Public comment meeting concerning public access to federally supported R&D data, National Academy of Sciences, May 16-17, 2013

- Find, understand, get, and (re)use data
- Robust metadata -- Complete, broad, understandable
  - Including information on provenance, methods, i.e. tools used, aggregation, processing and analytic techniques & versions, workflows, etc.
- Accessible, persistent, open institutional repositories of:
  - Data (input), *code* (processing & analysis), and publications (results)
- Persistent links relating:
  - Data, code, and publications
- Assure long-term access, transparency & "*really reproducible research*"
  - See especially Victoria Stodden's presentation "Why Public Access to Data Is So Important"